Multiple Regression Model

INSR 260, Spring 2009
Bob Stine
Overview

- Multiple Regression Model (MRM)
- Estimators, terminology (similar to SRM)
- Assumptions (new plots)
- Inference (new test)
- Prediction (similar to SRM)
- Examples (from Bowerman, Ch 4)
  - Fuel consumption
  - Sales management
Multiple Regression Model

- Equation has \( k \) explanatory variables
  - Mean: \( E(Y|X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mu_{Y|X} \)
  - Observations: \( y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \)

- Assumptions (as in SRM)
  - Independent observations
  - Equal variance \( \sigma^2 \)
  - Normal distribution around “line”
    \( y_i \sim N(\mu_{Y|X}, \sigma^2) \quad \varepsilon_i \sim N(0, \sigma^2) \)

- \( k+2 \) parameters identify model
  \( \beta_0, \beta_1, \ldots, \beta_k, \sigma^2 \)
Least Squares

Criterion
- Find estimates that minimize sum of squared deviations
  \[ \min_a \sum (y_i - a_0 - a_1 x_{i1} - \ldots - a_k x_{ik})^2 \]

Fitted values, residuals
- Fitted values (on the line) \( \hat{y} = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik} \)
- Residual deviations \( e = y - \hat{y} \)

Standard error of regression (estimate of \( \sigma^2 \))
- \( s^2 = \sum e_i^2 / (n-k-1) \)
- degrees of freedom
- RMSE = square root of \( s^2 \)
Goodness of Fit

- **R-squared statistic**
  - Square of correlation between $Y$ and $\hat{Y}$
  - Percentage of “explained” variation
  - Always increases as variables are added to equation

$$R^2 = \frac{\text{Explained SS}}{\text{Total SS}}$$

- **Adjusted R-squared**
  - Will not increase unless $s^2$ gets smaller
  - Difference from $R^2$ increases as $k$ increases

$$\overline{R^2} = 1 - \frac{s^2}{\text{var}(y)}$$
Checking Assumptions

- Scatterplots of $Y$ on $X_1$, $Y$ on $X_2$
- Data for fuel consumption ($n = 8$)

Scatterplot matrix

$y =$ weekly natural gas consumption
$X_1 =$ average temperature
$X_2 =$ chill index (wind, clouds, temp)

<table>
<thead>
<tr>
<th></th>
<th>FUELCONS</th>
<th>TEMP</th>
<th>CHILL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUELCONS</td>
<td>1.0000</td>
<td>-0.9484</td>
<td>0.8706</td>
</tr>
<tr>
<td>TEMP</td>
<td>-0.9484</td>
<td>1.0000</td>
<td>-0.7182</td>
</tr>
<tr>
<td>CHILL</td>
<td>0.8706</td>
<td>-0.7182</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Partial vs Marginal

### SRM

**Bivariate Fit of FUELCONS By TEMP**

\[ FUELCONS = 15.837857 - 0.1279217 \times TEMP \]

**Parameter Estimates**

| Term   | Estimate  | Std Error | t Ratio | Prob>|t| |
|--------|-----------|-----------|---------|------|
| Intercept | 15.837857 | 0.801773  | 19.75   | <.0001* |
| TEMP   | -0.127922 | 0.017457  | -7.33   | 0.0003* |

**Summary of Fit**

- Rsquare: 0.97363
- Rsquare Adj: 0.963081
- Root Mean Square Error: 0.367078
- Mean of Response: 10.2125
- Observations (or Sum Wgts): 8

### MRM

**Bivariate Fit of FUELCONS By CHILL**

\[ FUELCONS = 7.8494238 + 0.1835399 \times CHILL \]

**Parameter Estimates**

| Term   | Estimate  | Std Error | t Ratio | Prob>|t| |
|--------|-----------|-----------|---------|------|
| Intercept | 7.8494238 | 0.652654  | 12.03   | <.0001* |
| CHILL   | 0.1835399 | 0.042339  | 4.34    | 0.0049* |

**Summary of Fit**

- Rsquare: 0.97363
- Rsquare Adj: 0.963081
- Root Mean Square Error: 0.367078
- Mean of Response: 10.2125
- Observations (or Sum Wgts): 8

**Parameter Estimates**

| Term   | Estimate  | Std Error | t Ratio | Prob>|t| |
|--------|-----------|-----------|---------|------|
| Intercept | 13.108737 | 0.855698  | 15.32   | <.0001* |
| TEMP   | -0.090014 | 0.014077  | -6.39   | 0.0014* |
| CHILL  | 0.082495  | 0.022003  | 3.75    | 0.0133* |

Slopes differ
More Diagnostics

- Overall plots (MRM version of SRM scatterplots)

- Leverage plots (partial regression plots)
  - Simple regression view of MR slope, one for each slope
Inference

wives error of the slope is affected by correlation among explanatory variables.

- Variance inflation factor (Chap 5)
  \[
  \text{Var}(\text{slope in MRM}) \approx \text{Var}(\text{slope in SRM}) \times \text{VIF}
  \]

- Three equivalent methods for each estimated slope and the intercept
  - Confidence interval
  - t-statistic
  - p-value

![Parameter Estimates Table]

| Term  | Estimate | Std Error | t Ratio | Prob>|t| | Lower 95% | Upper 95% | VIF |
|-------|----------|-----------|---------|-------|----------|-----------|-----|
| Intercept | 13.108737 | 0.855698 | 15.32 | <.0001* | 10.909095 | 15.308379 | . |
| TEMP   | -0.090014 | 0.014077 | -6.39 | 0.0014* | -0.126201 | -0.053827 | 2.07 |
| CHILL  | 0.082495  | 0.022003 | 3.75  | 0.0133* | 0.0259356 | 0.1390543 | 2.07 |
Overall F Test

- Test both slopes simultaneously
  - $H_0: \beta_1 = \beta_2 = 0$
  - Ratio of variance explained to remaining variation
- Test of the size of $R^2$ statistic

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>24.875018</td>
<td>12.4375</td>
<td>92.3031</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>0.673732</td>
<td>0.1347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>7</td>
<td>25.548750</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Prediction

No simple plot
- Extrapolation effect is more subtle

Software is needed to identify extrapolation
- Options in Fit Model to save various standard errors as well as prediction and confidence intervals
- Add an extra row (before fitting) to get JMP to predict a new case

<table>
<thead>
<tr>
<th>FUELCONS</th>
<th>TEMP</th>
<th>CHILL</th>
<th>Pred Formula FUELCONS</th>
<th>StdErr Pred FUELCONS</th>
<th>Lower 95% Mean FUELCONS</th>
<th>Upper 95% Mean FUELCONS</th>
<th>StdErr Indiv FUELCONS</th>
<th>Lower 95% Indiv FUELCONS</th>
<th>Upper 95% Indiv FUELCONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.4</td>
<td>28</td>
<td>18</td>
<td>12.07</td>
<td>0.21</td>
<td>11.54</td>
<td>12.61</td>
<td>0.42</td>
<td>10.99</td>
</tr>
<tr>
<td>2</td>
<td>11.7</td>
<td>28</td>
<td>14</td>
<td>11.74</td>
<td>0.25</td>
<td>11.11</td>
<td>12.37</td>
<td>0.44</td>
<td>10.61</td>
</tr>
<tr>
<td>3</td>
<td>12.4</td>
<td>32.5</td>
<td>24</td>
<td>12.16</td>
<td>0.21</td>
<td>11.61</td>
<td>12.71</td>
<td>0.43</td>
<td>11.07</td>
</tr>
<tr>
<td>4</td>
<td>10.8</td>
<td>39</td>
<td>22</td>
<td>11.41</td>
<td>0.20</td>
<td>10.89</td>
<td>11.94</td>
<td>0.42</td>
<td>10.33</td>
</tr>
<tr>
<td>5</td>
<td>9.4</td>
<td>45.9</td>
<td>8</td>
<td>9.64</td>
<td>0.16</td>
<td>9.23</td>
<td>10.04</td>
<td>0.40</td>
<td>8.61</td>
</tr>
<tr>
<td>6</td>
<td>9.5</td>
<td>57.8</td>
<td>16</td>
<td>9.23</td>
<td>0.28</td>
<td>8.50</td>
<td>9.95</td>
<td>0.46</td>
<td>8.04</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>58.1</td>
<td>1</td>
<td>7.96</td>
<td>0.22</td>
<td>7.39</td>
<td>8.54</td>
<td>0.43</td>
<td>6.86</td>
</tr>
<tr>
<td>8</td>
<td>7.5</td>
<td>62.5</td>
<td>0</td>
<td>7.48</td>
<td>0.24</td>
<td>6.86</td>
<td>8.11</td>
<td>0.44</td>
<td>6.35</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>8</td>
<td>8</td>
<td>7.47</td>
<td>0.33</td>
<td>6.63</td>
<td>8.31</td>
<td>0.49</td>
<td>6.21</td>
</tr>
</tbody>
</table>
Sales Example

Question
- Evaluation of sales representatives
- Response is annual company sales in territory $y$ measured in thousands of units
- Data are a sample for $n = 25$ sales representatives

Several explanatory variables
- Time (months) with the company
- Total sales of company and rivals in territory (potential)
- Advertising expenditure in territory
- Company’s market share in prior four years
- Change in company’s market share
Initial Graphical Analysis

Scatterplot Matrix

Correlations

<table>
<thead>
<tr>
<th></th>
<th>Sales</th>
<th>Time</th>
<th>MktPoten</th>
<th>Adver</th>
<th>MktShare</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>1.0000</td>
<td>0.6229</td>
<td>0.5978</td>
<td>0.5962</td>
<td>0.4835</td>
<td>0.4892</td>
</tr>
<tr>
<td>Time</td>
<td>0.6229</td>
<td>1.0000</td>
<td>0.4540</td>
<td>0.2492</td>
<td>0.1062</td>
<td>0.2515</td>
</tr>
<tr>
<td>MktPoten</td>
<td>0.5978</td>
<td>0.4540</td>
<td>1.0000</td>
<td>0.1741</td>
<td>-0.2107</td>
<td>0.2683</td>
</tr>
<tr>
<td>Adver</td>
<td>0.5962</td>
<td>0.2492</td>
<td>0.1741</td>
<td>1.0000</td>
<td>0.2645</td>
<td>0.3765</td>
</tr>
<tr>
<td>MktShare</td>
<td>0.4835</td>
<td>0.1062</td>
<td>-0.2107</td>
<td>0.2645</td>
<td>1.0000</td>
<td>0.0855</td>
</tr>
<tr>
<td>Change</td>
<td>0.4892</td>
<td>0.2515</td>
<td>0.2683</td>
<td>0.3765</td>
<td>0.0855</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Multiple Regression

Overall fit

Residual by Predicted Plot

Actual by Predicted Plot

Leverage plots

Leverage Plot

Leverage Plot

MktPoten Leverage, P<.0001

Time Leverage, P=0.0065

Sales Leverage

Residuals

Sales Actual

Sales Predicted

Sales Leverage

Residuals

Sales Actual

Sales Predicted

Sales Leverage

Residuals

Sales Actual

Sales Predicted
Model Summary

Overall fit

Summary of Fit

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
<td>0.915009</td>
</tr>
<tr>
<td>RSquare Adj</td>
<td>0.892643</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>430.2319</td>
</tr>
<tr>
<td>Mean of Response</td>
<td>3374.568</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
<td>25</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>37862659</td>
<td>7572532</td>
<td>40.9106</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>3516890</td>
<td>185099</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>24</td>
<td>41379549</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual estimates

Interpretation of these estimates?
Why linear? Implications of model are very strong.

Parameter Estimates

| Term      | Estimate  | Std Error | t Ratio | Prob>|t| |
|-----------|-----------|-----------|---------|------|
| Intercept | -1113.788 | 419.8869  | -2.65   | 0.0157*|
| Time      | 3.6121012 | 1.1817    | 3.06    | 0.0065*|
| MktPoten  | 0.0420881 | 0.006731  | 6.25    | <.0001*|
| Adver     | 0.1288568 | 0.037036  | 3.48    | 0.0025*|
| MktShare  | 256.95554 | 39.13607  | 6.57    | <.0001*|
| Change    | 324.53345 | 157.2831  | 2.06    | 0.0530 |
Prediction

Conditions for another rep (not one of these 25)

- Sales were 3082
- Time with company 85.42
- Market potential 35,182.73
- Advertising 7,281.65
- Market share 9.64
- Change in share 0.28

Prediction results

- Plug values for explanatory variables into equation
- Prediction \( \hat{y} = 4182 \)
- Confidence interval for mean 3884.9 to 4478.6
- Prediction interval for rep 3233.6 to 5129.9
- Benchmarking implication: How is this rep doing?
Summary

- Multiple Regression Model (MRM)
- Estimators: partial (MRM) vs marginal (SRM)
- Assumptions: leverage plots
- Inference: F-test
- Prediction: Software